ROOM AND LOW TEMPERATURE PERFORMANCE OF HIGH-SPEED NEURAL NETWORK CIRCUITS

Tuan Duong, Taher Daud, Anil Thakoor,
Center for Space Microelectronics Technology
Jet Propulsion Laboratory
California Institute of Technology, Pasadena, CA 91109
and
Bruce Lee
Irvine Sensors Corporation, Costa Mesa, CA 92626

ABSTRACT

We have described a challenging neural-network hardware implementation that would mate a 64x64-pixel infrared sensor directly on to a 3-dimensionally packaged set of neural processing chips with parallel input for high speed image processing. We now describe two enhanced schemes. In one the synapse (analog multiply device) resolution has been increased from 7- to 8-bit without increasing the silicon area but requiring low temperature (77K) operation. In other an innovative chip is incorporated in place of the mated infrared sensor array that permits detachment of the IR sensor without compromising the input speed. This permits room temperature operation of the processor for a projected image processing speed of a tera-operations per second (TOPS). We compare the 7- and 8-bit synapse architectures and provide the test results of the operation (linearity and speed) both at low and room temperatures. Non linearity effects and range compression issues are also discussed.

INTRODUCTION

Artificial neural networks, derived from their biological counterparts with their inherent massive interconnectivity and parallel processing architecture, are specially suitable for image and signal processing that requires feature classification/object recognition, global optimization, and adaptive control [1]. When implemented in fully parallel electronic hardware, they provide orders of magnitude speed enhancement [2]. VLSI-implemented neural network chips that are wired with nearly full parallelism have been shown to reduce processing time by orders of magnitude and are useful in a variety of applications [2,3]. Basic building blocks of the neural network architecture are the processing elements called "neurons" implemented as nonlinear operational amplifiers with a sigmoidal transfer function, interconnected through weighted connections called

"synapses" implemented using circuitry for digital weight storage and analog multiply (between input and stored weight value) functions [4,5]. Taking cue from biology of the massive parallelism for image data input from retinas to cerebral cortex, the overall effectiveness of the neural network hardware was enhanced via an innovative architecture for a tight coupling between a sensor and the processing chips as was reported in the last meeting [6].

This architecture consisted of mating a 64 x 64 infrared (IR) image sensor to a stack of 64 neural net ICs directly along edges, each with different stored weights. A variety of image processing tasks could be performed in parallel at extremely high speeds and in a highly compact package (≈1 inch cube) [7]. The simultaneous requirements dictated by such an application on the integrated neuroprocessing cube were:

(a) Cold temperature operation, IC stack being mated to the infrared (IR) imager required to operate at ~90°K.
(b) Low power dissipation of ≈2 watts because of the need to maintain cold temperatures.
(c) High speed operation approaching 1000 frames per second, which translated into a 4 MHz pixel image processing rate, and hence a <250 nanoseconds signal processing speed.

These requirements had led to the design of low power digital-analog hybrid circuits for implementation of the VLSI neural network ICs. Use of analog circuitry for signal flow processing enabled a very compact, low power neural network realization [3]. However, digital circuitry was also judiciously used for weight storage at synapses, utilizing the static random access memory (SRAM) concept [4]. The 7-bit storage register for the synapse was described [6, 8]. However, before its implementation, it was deemed necessary to update the synapse circuit to have 8 bits of resolution. This was accomplished not by adding additional similar circuitry for the 8th bit which would have required nearly doubling of the circuitry on the chip and hence doubling the silicon real estate, but by using an innovative concept of splitting the circuit into two, one part acting as a vernier for the other. This significant design innovation resulted in a marked saving of chip area as well as power. However, a trade-off was the reduction in signal dynamic range because of the anticipated nonlinearity of the characteristics which was also confirmed after chip testing.

The architecture, shown in Figure 1 (A), was dedicated to one IR imager array mated to the processor and hence the requirement for low temperature (~77K) operation of the neural net chips. While this implementation is still in progress at the time of this writing, in a parallel development it was decided to use an identical processor cube to further enhance the architecture by detaching the IR sensor array from the neural processing module (NPM). Therefore, a new chip (to be mated to the processor) was designed to obtain image data from any type and size imager (say a 256x256-pixel array)

by rastering a 64x64 window along the image and feeding it to the processor with full parallelism and without compromising the speed. This revised architecture is shown in Figure 1 (B). Now it was no longer necessary for this processor cube to operate at 77K. However, since the synapse design was originally optimized for low temperature operation, it was imperative that its room temperature performance be satisfactory for the new architecture to be viable.

The paper describes overall architectures of the two approaches and focuses on the extension of the synapse resolution from 7- to 8-bit. Test results of the VLSI-implemented NPM chip at room and cold temperatures are presented. A brief discussion of the impact of 8-bit synapse design on the performance tradeoff is also given.

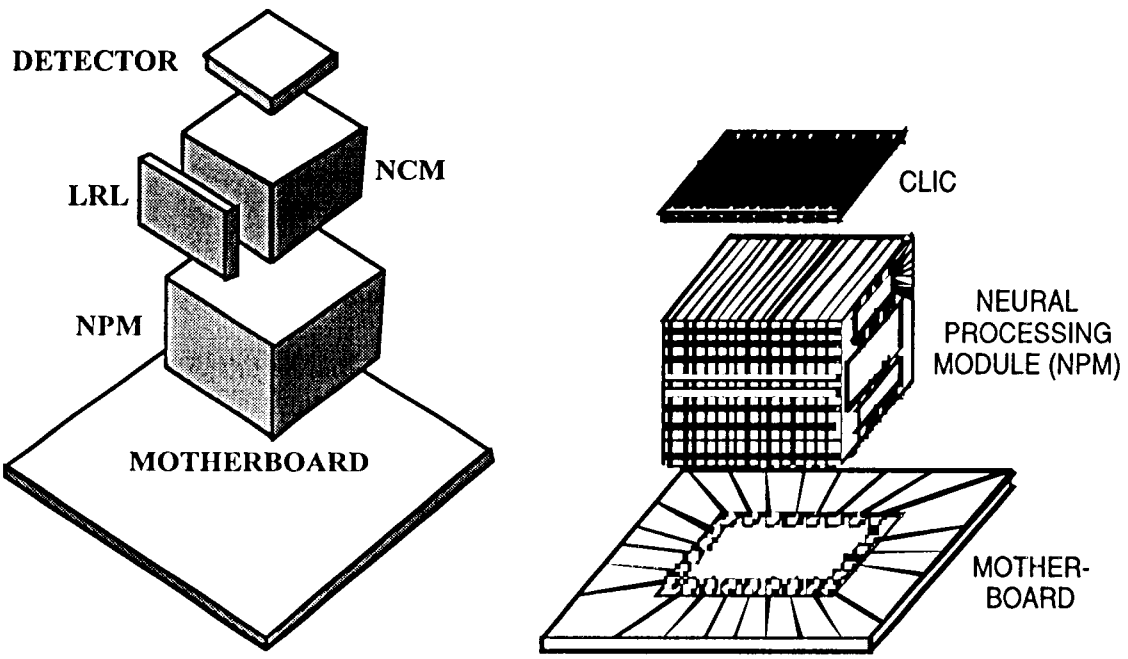


Figure 1. The schematic views of the two 3D-stacked implementations. (A) The 3D Artificial Neural Network (3DANN) that has an attached 64x64 IR detector array. The lateral resistive layer (LRL) chip performs the gaussian blurring or deblurring of the image depending on its programmed resistance values (10 kilo- to 10 meg-ohms). (B) The detector and neural conditioning module (NCM) have been replaced by a column loading input chip (CLIC) that would raster a 64x64 window of a larger IR, visible, etc. images. The NPM cube is identical in both implementations. Because of these changes, however, the two motherboard designs with metal lines for bonding to the pads at the identical packages are slightly different.

MODIFIED NEURAL NETWORK ARCHITECTURE

The stacked architecture promises a practical realization of three dimensional electronic circuitry, offering unprecedented computational power in such a compact package [7]. Figure 1 illustrates the two approaches using the same neural processing module (NPM). Sixty-four thinned VLSI-implemented NPM chips will be stacked to form a three-dimensional "sugarcube". Using the bump bonding technique, either a 64x64 IR sensor array as per the original architecture or a newly designed Column Loading Input Chip (CLIC) will be mated to the IC stack as per the second scheme. Each of the 64x64 output in the CLIC is brought out to a respective pad maintaining the same geometrical constraints and is directly bump-bonded to connect to the respective inputs of the NPM chips. Communication amongst the stacked ICs is made possible by providing meta bus lines running across side planes.

Each NPM chip in the stack is identical and essentially consists of 64 input lines, 64 output lines, and a 64x64 synaptic array where each synapse provides a product of the incoming analog signal with its digital stored weight as an analog output current. A photograph of the NPM chip is shown in Figure 2. The individual outputs from 64 synapses along each of the 64 columns are summed on a chip and then for all the 64 chips providing the final 64 output currents. Each synapse weight is electrically and

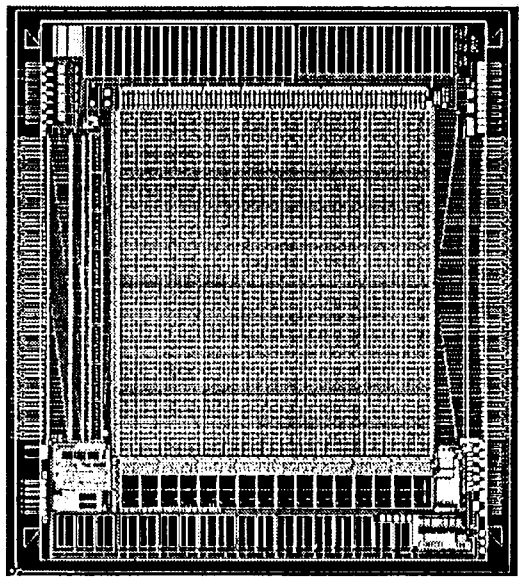


Figure 2. A photograph of the neural processing module (NPM) chip showing the 64x64 array of 8-bit synapses along with analog incoming circuitry, and digital weight loading and other control circuitry.

individually programmable under software control. This architecture allows flexibility in the computational functionality. For instance, each of the 64 neural networks could look at the same image window and convolve with different weight templates, or the entire image could be divided into a set of 64x64 windows stored as weights with one incoming pattern as a parallel convolver.

Offering high density and massively parallel "focal plane" processing capability, these smart packages would be demonstrated for the fast frame seeker function involving real time (64x64-pixel, at 1000 frames/s) image acquisition, recognition, and tracking. Similarly, the digital weight storage for the chips is provided in parallel on all the 64 chips by 64 input lines at a rate of 2 gigabits per second. Its compact footprint would further complement the 64-chip cube consuming only about 2.5 watts of power during its data processing operation [6]. Each column weight values on the 64 chips form a 64x64 template and the incoming 64x64 image is convolved with all 64 templates stored on the NPM cube within a matter of 250 nanoseconds, giving 64 convolution outputs, hence the unprecedented speed of about one teraoperations per second. Here an operation is defined as one analog multiply-accumulate equivalent.

ELECTRONIC DEVICE DESIGNS

NPM chips incorporating an array of 64x64 synapses along with other switching, weight loading, and control circuitry were fabricated after innovatively modifying the synapse design to increase its bit-resolution from 7 to 8. Our synapse designs are based on a static random access memory (SRAM) architecture for digital loading of weights combined with its digital-to-analog multiplication with an analog input signal giving an analog current output. The chips were fabricated with a 64x64 array of a $120x120\mu m^2$ unit cell in a 1.2μm CMOS fabrication process.

7-Bit Synapse Design

The 7-bit synapse circuit [6], shown in Fig. 3 (A), consists of a voltage-to-current input, a 7-bit multiplying digital to analog converter (MDAC), and a 7-bit digital memory. This type of synapse, with its on-chip storage of digital weights, allows a very simple digital interface (as opposed to the need for refresh circuitry to update volatile analog storage of weights) and has been successfully incorporated into a number of our implementations [3]. The current realization utilizes single transistor current mirrors rather than the cascode current mirrors of previous designs. This difference results in higher speed and a more compact design at the cost of a possible decrease in circuit robustness.

Operation of a synapse cell is as follows. An input transistor, biased in the linear region (Vdrain < Vgate - Vt), converts an input voltage (V_{in}) applied to its gate into a drain current (I_{in}) which is almost linearly proportional to V_{in}. This input current is then multiplied by the stored digital word (weight) to produce the desired output current (I_{out}). Multiplication is accomplished by conditionally scaling the input current I_{in} by a series of current mirror transistors. For each current mirror, a pass transistor controlled by one bit of the digital word conditionally allows current to be placed on a common summation line. The bits in the digital word from LSB to MSB are connected to 1, 2, 4, 8, 16, and 32 current mirror transistors respectively so that the input current is scaled by the appropriate amount. The resulting summation current is unipolar. However a current steering differential transistor pair, controlled by the seventh bit of the digital word, determines the direction of the output current, such that two-quadrant multiplication is accomplished (-63 to +63 levels). The 7-bit digital memory consisting of 7 static latches provides programmable, nonvolatile weight storage and is randomly accessible. One input transistor circuit is coupled through current mirrors to all the synapses along one column in the input synapse matrix (or row for output synapse matrix) because the current I_{in} is required as input to all the synapses in that column (rows for the output).

8-Bit Synapse Design

The 8-bit synapse circuit shown in Fig. 3 (B) is obviously a departure from the normal extension of the 7-bit synapse circuit. As one can visualize from Figure 3 (A), in an ordinary extension to 8-bit synapse circuit the bits in the digital word from LSB to

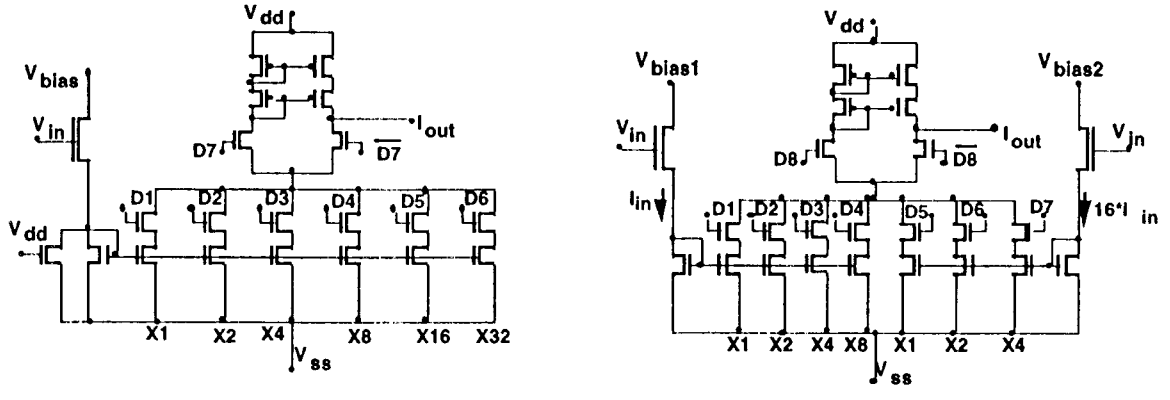


Figure 3 Digital-analog hybrid synapse cells for multiply-accumulate: (A) Circuit diagram of a 7-bit (±64 levels) synapse cell; (B) Circuit diagram of an 8-bit (±128 levels) cell. Current summing in an analog domain is obtained from multiple synapse output nodes by connecting them to a common wire.

MSB would have to be connected to 1, 2, 4, 8, 16, 32 and 64 current mirror transistors, respectively so that the input current is scaled by the appropriate amount, thereby requiring additional 64 current mirror circuits for a total of 127 (63+64), nearly doubling the transistor count.

Instead, as shown in Fig. 3 (B), two input transistors biased in the linear region are used to convert the common input voltage Vin into two input current values that have a 16:1 ratio using two appropriate bias currents given by V_{bias2} and V_{bias1} respectively. The bits in the digital word for the higher current circuit (right side) are accordingly connected to 1, 2, and 4 current mirror transistors providing steps of 16*Iin. Complementing this, the lower current (left) side has 1, 2, 4, and 8 such current mirror circuits providing the vernier type steps of Iin filling in the required 15*Iin steps. Thus the complete range from 0 to +127*Iin is taken care of by a combination of the two sets of circuits with D1 through D7 latches. The current steering circuit is identical to the one for the 7-bit synapse with a latch D8 for the 8th bit, providing the complete range ±127*Iin.

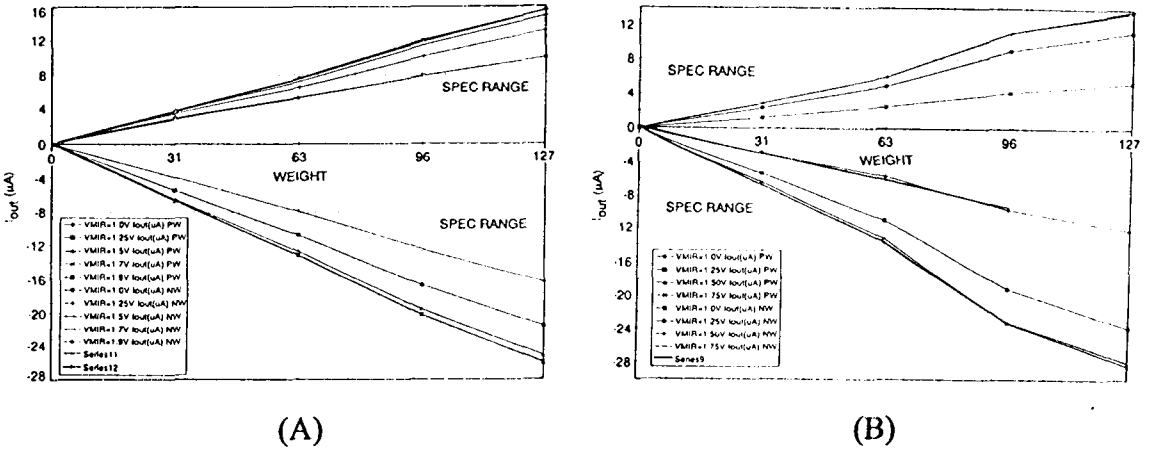
RESULTS

The circuits were simulated at 77 K temperature (using a PC version of PSPICE simulator specially suitable for low temperature circuit modeling) and the design was optimized. Because of the changes in the synapse design as a result of the increase in its dynamic range from 7 to 8-bit, interesting tradeoffs have been made. The chip test results have shown that the synapse performs the 8-bit weighting function at 77°K at high speed (150ns) but with good linearity only within a narrower range of bias voltage (±1 volt). On the contrary, the same design provides much better linearity at room temperature but with a speed penalty (300ns). Figure 4 (A) & (B) show the measurement results at room temperature and at 77 K, respectively. The readings for the negative weight values are folded over for compactness of the graph and show up as -ve output currents. The specification range for a linear output is restricted to an output current range of ±8 μA for a 1 volt input swing (2.5 to 3.5 volts). Further, measured data on power consumption of the NPM chip exceeds the designed number of ~2W for the 64-chip cube with 64x64 synapse-neuron circuit on a chip by about half a watt.

CONCLUSIONS

The high speed digital-analog synapse circuits have been designed and fabricated as neuroprocessing module chips to operate at 77 K, and tested successfully for high speed processing both at room and low temperatures. Nonlinear effects have been taken care of by restricting the range of input signals to 1 volt. A synapse multiply-accumulate

throughput speed of ~250 ns has been obtained which translates into about one teraoperations per second (orders of magnitude higher than that of conventional processing computers) for specialized high speeed image convolution functions. In addition, tests have shown that the total power consumption for the NPM with 64 chips would be about 2.5 watts under the worst operating conditions. These results project that when implemented as a 'sugar cube', it will perform the inner-product image convolution and template matching operation at extremely high speeds of a teraoperations per second. Such a powerful computing engine combined with a versatile SIMD machine (a CNAPS board) interfaced with a Pentium Pro machine is being targeted for field demonstration on IR and visible images of BMDO interest for object recognition, discrimination and tracking in real time.



(A) (B)

Figure 4. (A) Room temperature and (B) Low (77 K) temperature characteristics of an 8-bit synapse [output vs. digital weights] on the neural processing module (NPM) chip. Linearity is maintained within the specification range (V_{bias} = VMIR = ±1 volt).

ACKNOWLEDGMENTS

The research described herein was performed in part by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology and the Irvine Sensors Corporation, and was jointly sponsored by the Ballistic Missile Defense Organization (BMDO), and the National Aeronautics and Space Administration (NASA). The authors are thankful to S. Suddarth, S. Udamkesmalee, J. Carson, C. Saunders, M. Skow, L. Lome, and S.K. Khanna, for useful discussions. Technical help by T. Thomas is also appreciated.

REFERENCES

1. R.P. Lippmann, "An introduction to computing with neural nets," IEEE ASSP Magazine, April 1987, pp. 4-22.
2. S.P. Eberhardt, T. Daud, D.A. Kerns, T.X Brown, and A.P. Thakoor, "Competitive neural architecture for hardware solution to the assignment problem," Neural Networks, Vol. 4, no. 4, pp. 431-442, 1991.
3. S.P. Eberhardt, R. Tawel, T.X Brown, T. Daud, and A.P. Thakoor, "Analog VLSI neural networks: Implementation issues and examples in optimization and supervised learning," IEEE Trans. Industrial Electronics, Vol. 39, no. 6, pp. 552-564, 1992.
4. A. Moopenn, T. Duong, and A.P. Thakoor, "Digital-analog hybrid synapse chips for electronic neural networks," Proc. IEEE Neural Information Processing Systems Conference, Denver, 1989.
5. S. Eberhardt, T. Duong, and A. Thakoor, "Design of parallel hardware neural network systems from custom analog VLSI 'building block' chips," Proc. IEEE/International Joint Conference on Neural Networks, Washington, DC, 1989, vol. II, pp. 183-190.
6. T. Daud, T. Duong, S. Kemeny, M. Tran, and A. Thakoor, "Low temperature performance of high-speed neural network circuits," ECS Proceedings of the Symposium on Low Temperature Electronics and High Temperature Superconductivity; Eds: Claeys, C.L., et al., vol. 95-9, 1995, pp. 334-344.
7. J. Carson, "On-focal plane array feature extraction using a 3-D artificial neural network (3DANN)," Proc. SPIE, Vol. 1541, Infrared Sensors: Detectors, Electronics, and Signal Processing, Ed: T.S.J. Jayadev, Part I: pp. 141-144, Part II: pp. 227-231, 1991.
8. T. Duong, S. Kemeny, M. Tran, T. Daud, and A. Thakoor, "Low power analog neurosynapse chips for a 3-D 'sugarcube' neuroprocessor," Proceedings of the IEEE International Conference on Neural Networks, WCCI, Vol. III, 1994, Orlando, FL, pp. 1907-1911.